Load Balancing in Cellular Networks with User-in-the-loop: A Spatial Traffic Shaping Approach

Ziyang Wang, Rainer Schoenen,
Halim Yanikomeroglu, Marc St-Hilaire

Department of Systems and Computer Engineering
Carleton University, Ottawa, Ontario, Canada

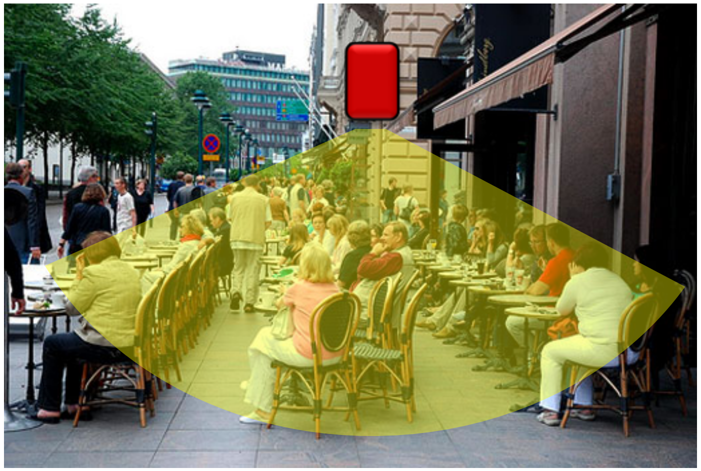
Carleton UNIVERSITY

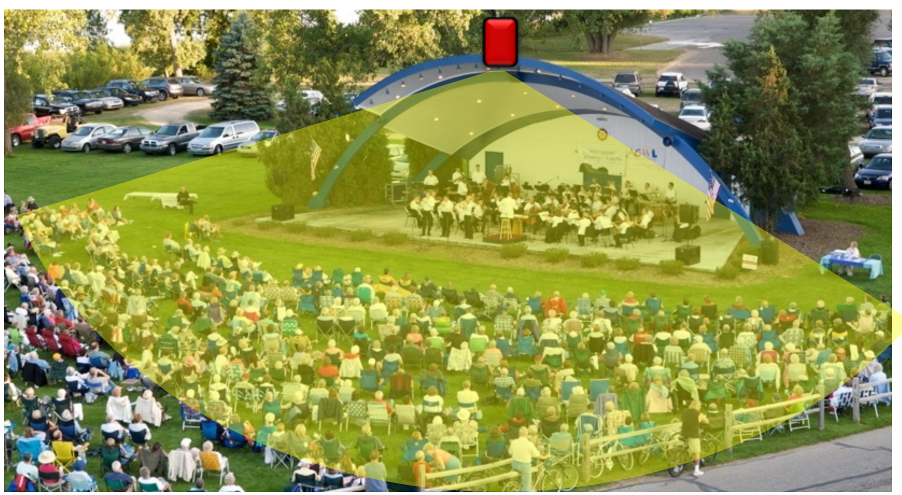
Sources of User Spatial Heterogeneity: 1) Self-Clustering

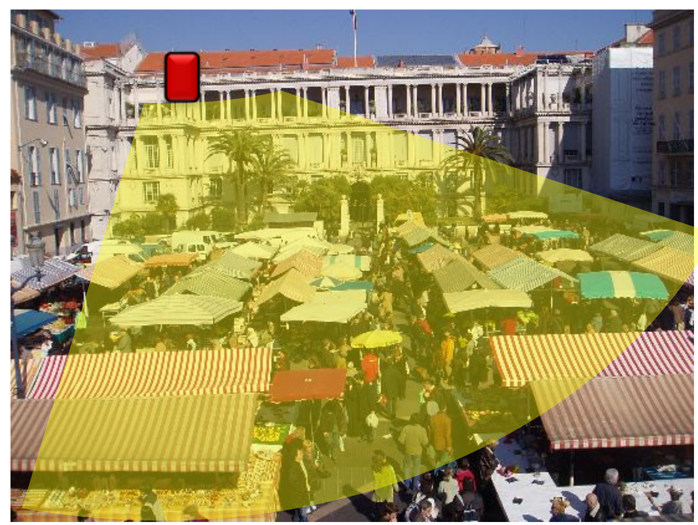
Sources of User Spatial Heterogeneity: 2) Urban Layout

Sources of User Spatial Heterogeneity: 3) Fixed Social Attractors

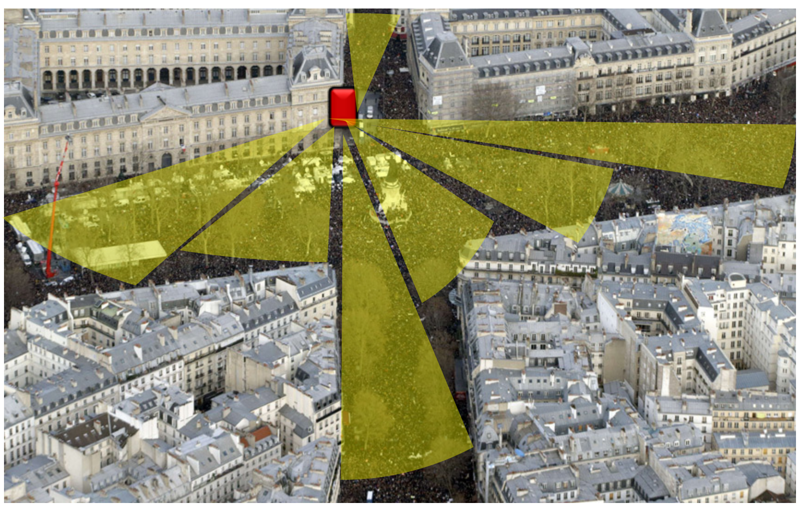
Small Cell Planning around Social Attractors
➔ Correlation between user clusters and AP locations

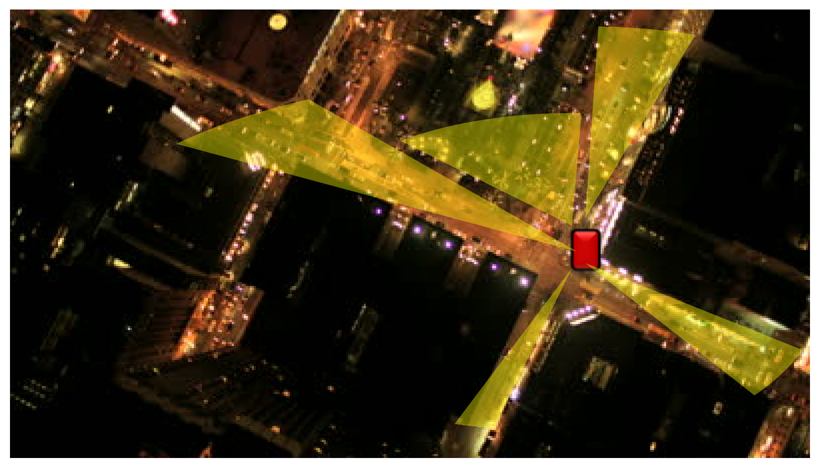






Small Cell Planning around Urban Layout
➔ Correlation between user layout and AP locations





Heterogeneity in Applications

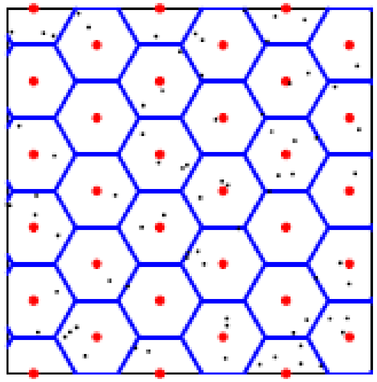
M2M Module	=	3 X	*
Wearable Device	=	6 X	*
Smartphone	=	37 X	*
Tablet	=	94 X	*
Laptop	=	119 X	*

* Monthly basic mobile phone data traffic.

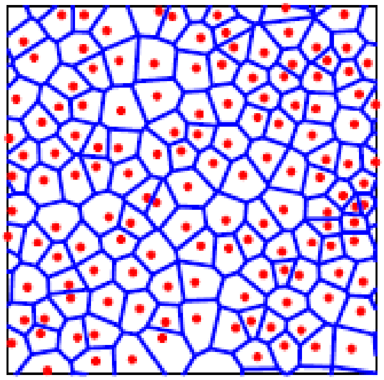
Source: Cisco VNI Mobile, 2015

HetNets: Heterogeneity in Supply (Access Points)

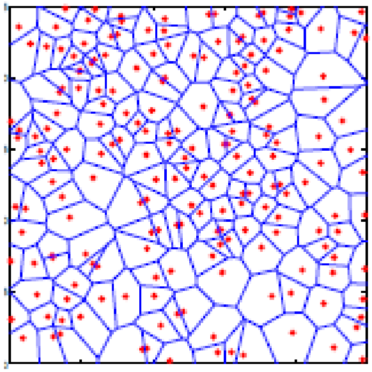
Locations of APs somewhere between a regular grid and total randomness:



Traditional grid model

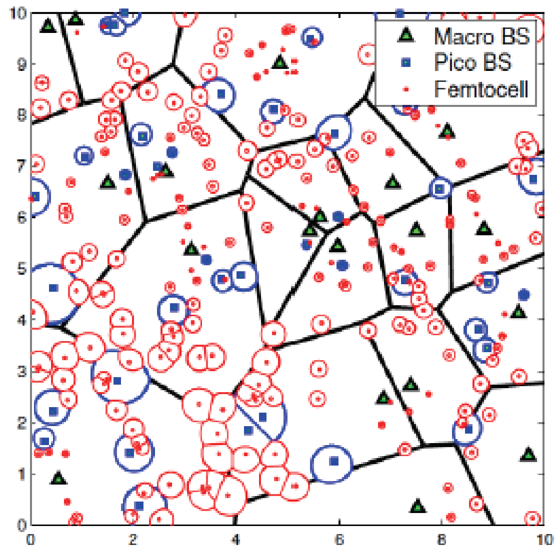


Actual 4G network today



Completely random BSs

Coexistence of several cells types with very different coverage ranges.



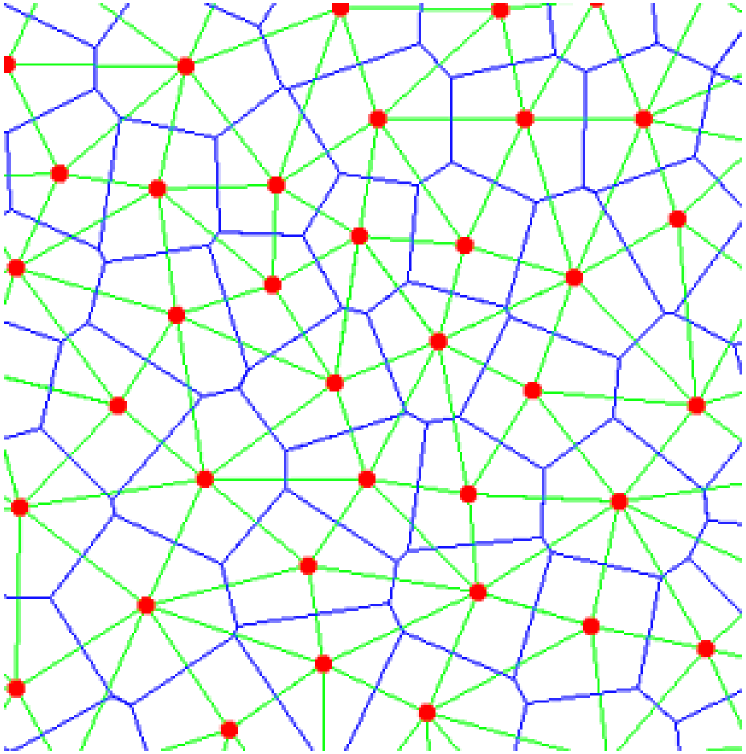
[U of Texas, Austin]

Measuring Spatial Heterogeneity: Step 1: Voronoi Tesselation

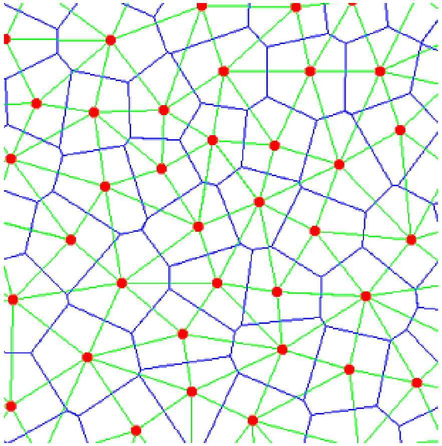
- For a set of Points, Find the Voronoi Partition: areas that are closer to their own point than any other point.

- Two points are “natural neighbours” if their Voronoi cells touch.

- Natural neighbours are connected by straight edges to form the Delaunay triangulation.



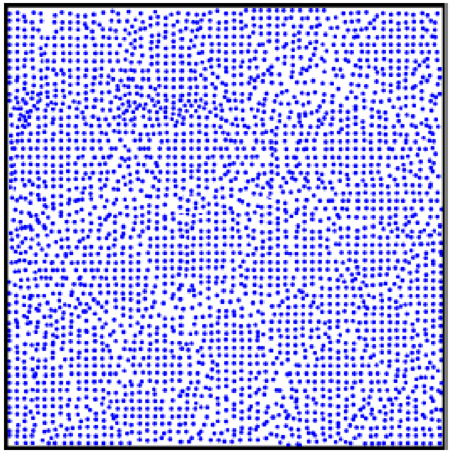
Measuring Spatial Heterogeneity: Step 2: Coefficient of Variation



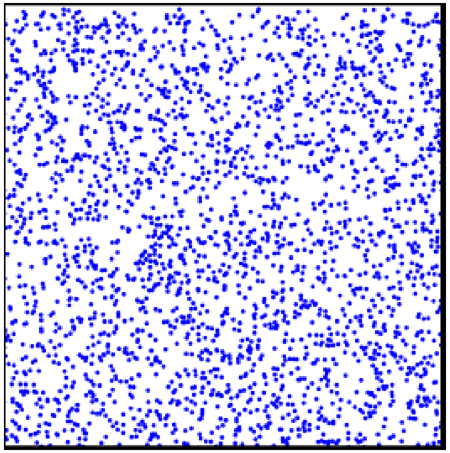
- Statistic: Coefficient of Variation:

  CoV{x} = std.dev.{x}/(mean{x} * K)   [K: a constant]
- We study two metrics (two “flavours”):
  - CoV of Voronoi Cell Areas (K=0.529)
  - CoV of Delaunay Cell Edge Lenghts (K= 0.492)
- CoV (either flavour) captures heterogeneity(dispersion/clustering) of any point process in one positive scalar value:

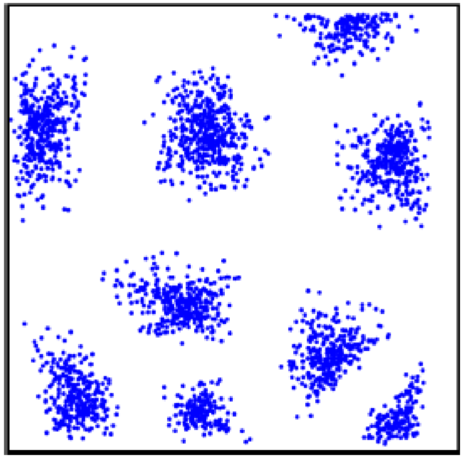
sub-Poissonian (e.g., repulsive): 0<CoV<1



Poisson Point Process: CoV=1

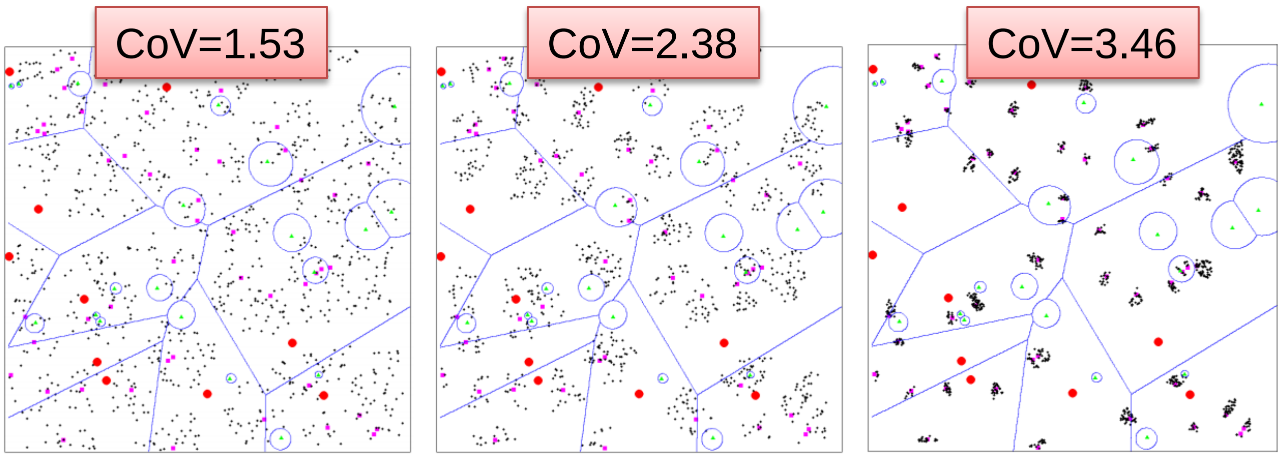


super-Poissonian (e.g. clustered): CoV>1



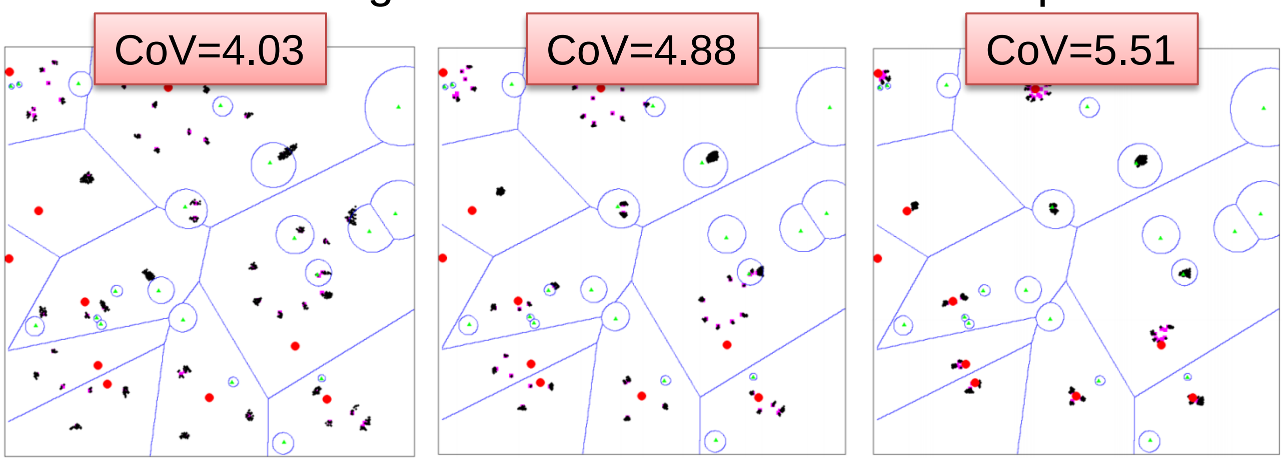
HetHetNets = HetNets + Heterogeneity in Demand (User Locations)

Users (black) self-clustering: clustering increases with beta



(b) $\alpha = 0$ and $\beta = 0.3$ (c) $\alpha = 0$ and $\beta = 0.6$ (d) $\alpha = 0$ and $\beta = 0.9$

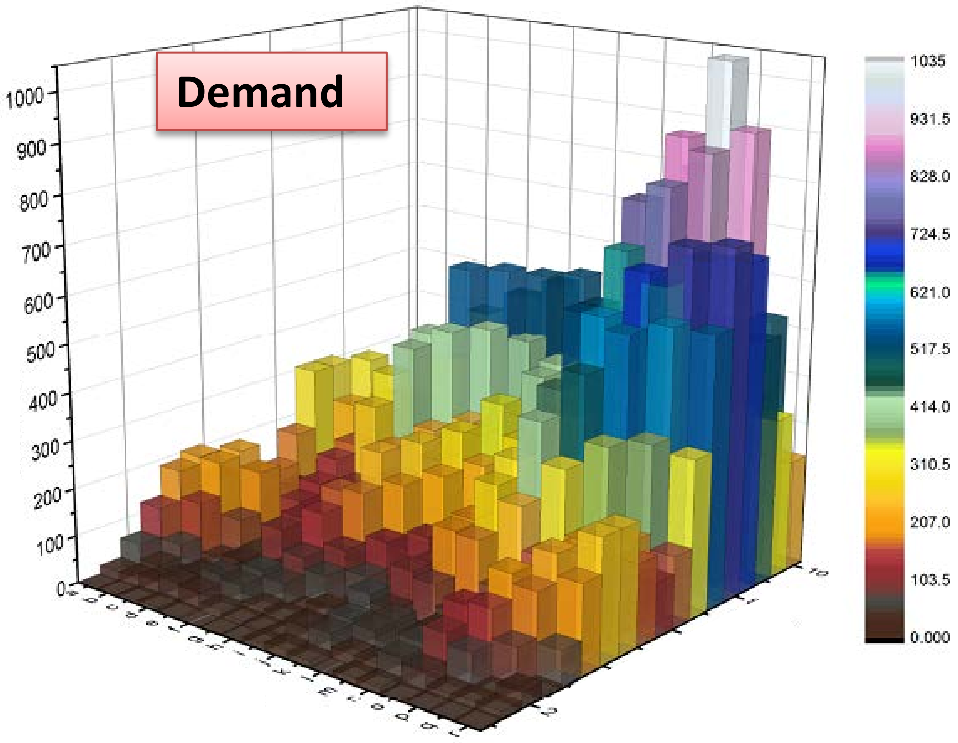
User clustering around APs: increases with alpha

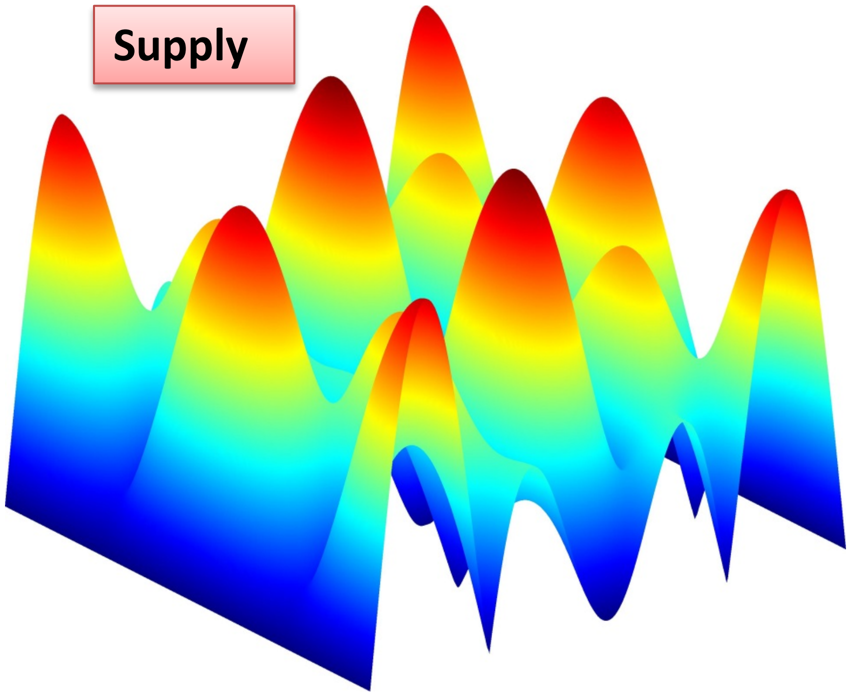


(h) $\alpha = 0.3$ and $\beta = 0.9$ (l) $\alpha = 0.6$ and $\beta = 0.9$ (p) $\alpha = 0.9$ and $\beta = 0.9$

[JSAC Oct 2015]

If Supply and Demand Do Not Match in Space and Time…





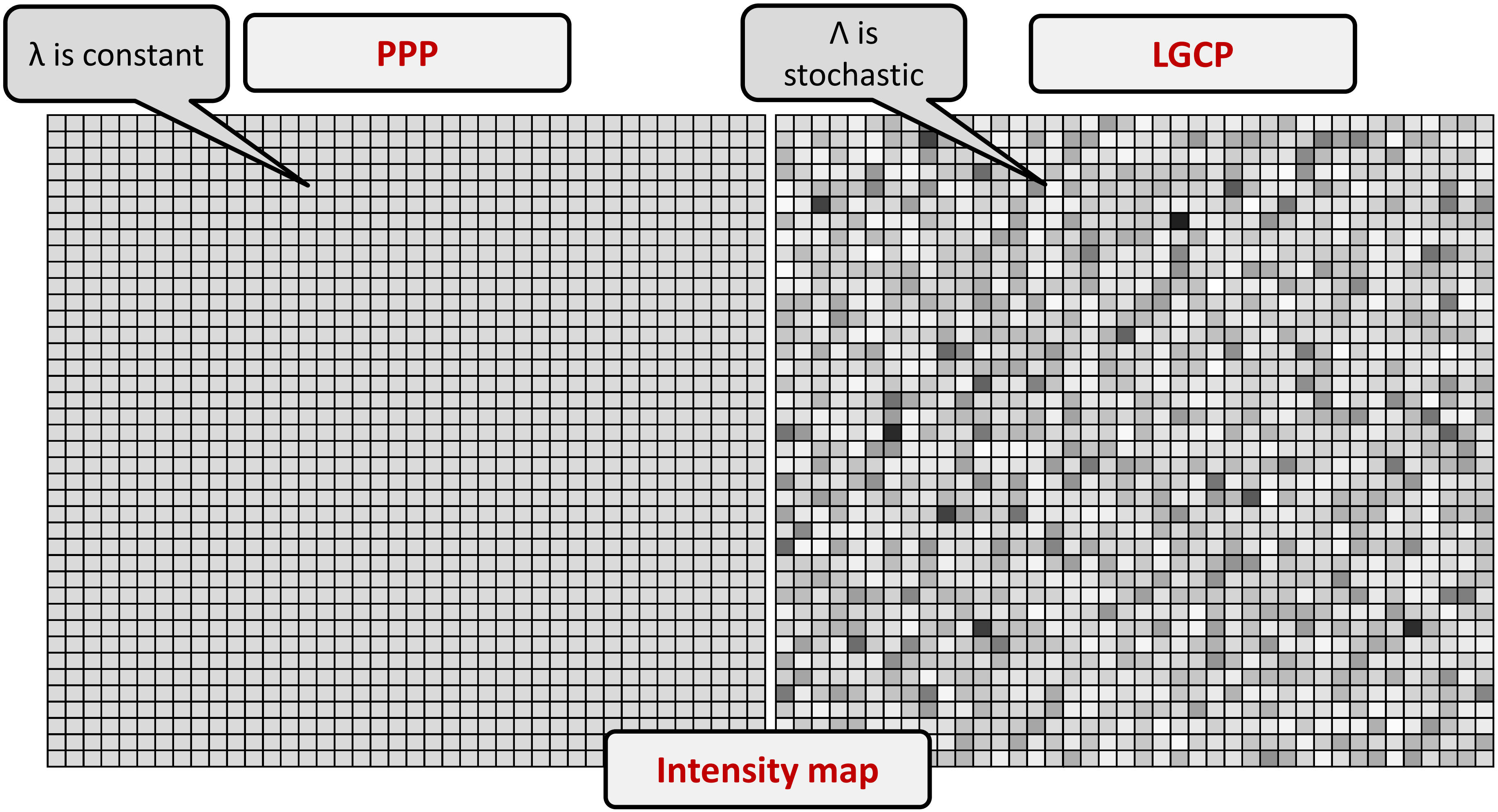
Can we store (in time) and/or transfer (in space) the supply?

If difficult, then more heterogeneous + more unpredictable → more problems

Log Gaussian Cox Process (LGCP)

- Cox process is a generalization of the PPP, also known as Doubly Stochastic Poisson Process.
- The intensity in Cox Λ is itself a stochastic process.
- In a PPP, for any bounded area B, the number of points in B is a Poisson number with mean $\lambda \cdot A_B$.
- In a Cox process, the number of points in B is a Poisson number with mean $\int_B \Lambda(s)ds$.
- A Cox process is a LGCP if $\Lambda(s) = \exp(Y(s))$, where $Y = \{Y(s): s \in R^2\}$ is a real valued Gaussian process.
- By changing the σ in Y, the LGCP generates a wide range of heterogeneities.

Realization of LGCP



User – BS Association

- User rate = (spectral efficiency) x (resource allocated)

- Max-SINR cell association: 2G, 3G, even 4G.

  Received SINR (spectral efficiency) is maximized, but results in load imbalance, especially in HetHetNets.

- Load-aware cell association: Balances the load, but sacrifices user spectral efficiency for better share of resources.

User – BS Association

- User rate = (spectral efficiency) x (resource allocated)

- Max-SINR cell association: 2G, 3G, even 4G.

  Received SINR (spectral efficiency) is maximized, but results in load imbalance, especially in HetHetNets.

- Load-aware cell association: Balances the load, but sacrifices user spectral efficiency for better share of resources.

- Can we simultaneously increase spectral efficiency and allocated resources for higher rates?

Carleton UNIVERSITY

User-in-the-Loop: Demand Shaping in Space and Time

ACCEPTED FROM OPEN CALL

IEEE Communications Magazine • February 2014

User-in-the-Loop: Spatial and Temporal Demand Shaping for Sustainable Wireless Networks

Rainer Schoenen and Halim Yanikomeroglu, Carleton University

wikipedia.org/wiki/user-in-the-loop

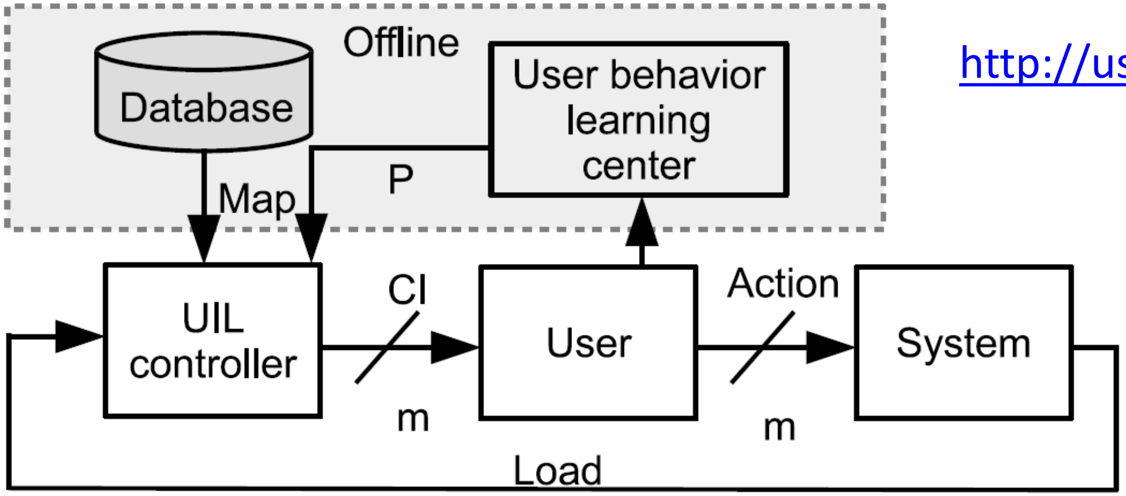
ABSTRACT

The demand for wireless access data rates is growing exponentially at a pace where supply cannot keep up with. Wireless resources (spectrum, time, space) are limited and shared, and transmission rates cannot be improved anymore solely with physical layer innovations. On the consumer side, flat rate type tariffs have established unnecessarily high expectations and often wasteful consumption. Dealing with congestion is unavoidable as a consequence of operating in a regime where demand is close to, equal to, or exceeding the supply. We can no longer assume that the current over-provisioning approach continues to be feasible.

ers come with cellular interfaces built in. In addition, user behavior is changing towards using more features and apps on their user terminal (UT). Many of them are not necessary or urgent at the moment, but users have built up habits and expectations that they must work everywhere with the same quality. Commercial stakeholders want users to consume more as long as the market is young and dynamic. As an inevitable consequence, this wireless traffic demand is growing faster than the supply side. Current analyses estimate that the almost 100 percent per year growth rate will continue for the next decade [2–5].

On the other side, scientists and engineers are working hard to push technology to the limit. Due to limitations by physics (like the Shannon

System Model



http://userintheloop.org

- **Map:** city map and spectral efficiency map
- **CI:** control information shown on users' terminal devices in the form of suggestions.
- **Action:** users can choose to comply with the suggestions or not
- **Load:** the load of each cell in the system
- **P:** the probability of each user to move to different locations. It is the output of an offline user behavior learning center. It could be formulated as $P_u(\text{distance } d, \text{QoS } q, \text{Incentive } i, \text{User Context } c)$. In this paper, we adopt the results of the previous research, which is the function of (*d*, *q*, *i*).

User Model and Resource Allocation

- Heterogeneous user spatial distribution (by LGCP)
- Heterogeneous user traffic class
  - Best effort (BE)
  - Guaranteed bit rate (GBR). Suppose a fixed rate *r* for all GBR users.
- The resources need for GBR users to reach rate *r* is $w_{ij} = \frac{r}{s_{ij}}$
  - s_{ij} is the spectral efficiency between user *i* and cell *j*
- A GBR user will get the exact amount of resources w_{ij} if

$$W_j - \sum_{i' \in U^g(i)} a_{i'j} w_{i'j} > w_{ij}$$

else, this GBR user is blocked, i.e., an outage occurs.

  - $U^g(i)$ is the set of all the existing GBR users in cell *j* when user *i* arrives to the system.

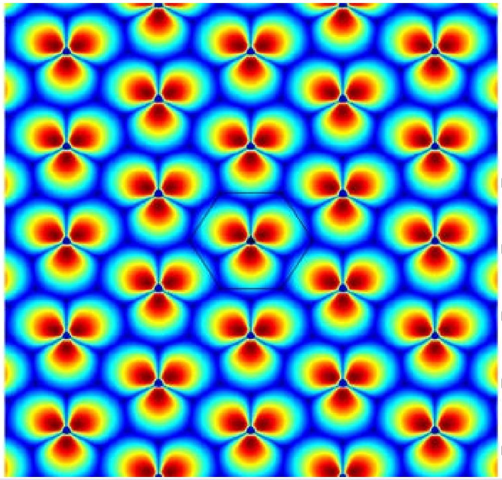
- The amount of resources allocated to a BE user *k* is

$$w_{kj} = \frac{W_j - \sum_{i \in U^g(k)} a_{ij} w_{ij}}{n_j^b(k) + 1}$$

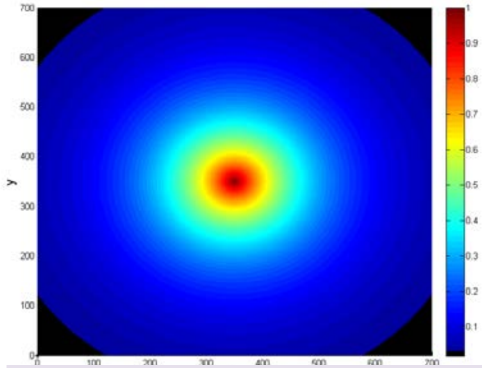
Utility Function

- The utility of a GBR is $U_{ij}(x,y) = p_i(x,y) \cdot s_j(x,y) \cdot (1 - \rho_j^b(i))$
  - $\rho_j^b(i) = \frac{\sum_{i' \in U^g(i)} a_{i'j} w_{i'j}}{W_j}$, the load factor of GBR users of cell *j* when user *i* arrives to the system
  - $p_i(x,y)$ is the probability of user *i* moving from the current location to (x, y)
  - $s_j(x,y)$ is the spectral efficiency map of cell *j*
- The utility of a BE user is $U_{kj}(x,y) = p_k(x,y) \cdot s_j(x,y) \cdot \frac{(1-\rho_j^b(k))}{n_j^b(k)+1}$
  - $n_j^b(k)$ is the number of BE users when user *k* arrives to the system
  - $\rho_j^b(k)$ is the load factor of the existing GBR users of cell *j* when BE user *k* arrives to the system
- ρ ϵ [0,1]. The difference between the utilities for GBR (guaranteed!) and BE comes from the fact that BE users can/must share the remaining capacity.
- The utility function generates a three dimensional matrix; the first dimension is the cell index, and the other two dimensions are the coordinates of the map.

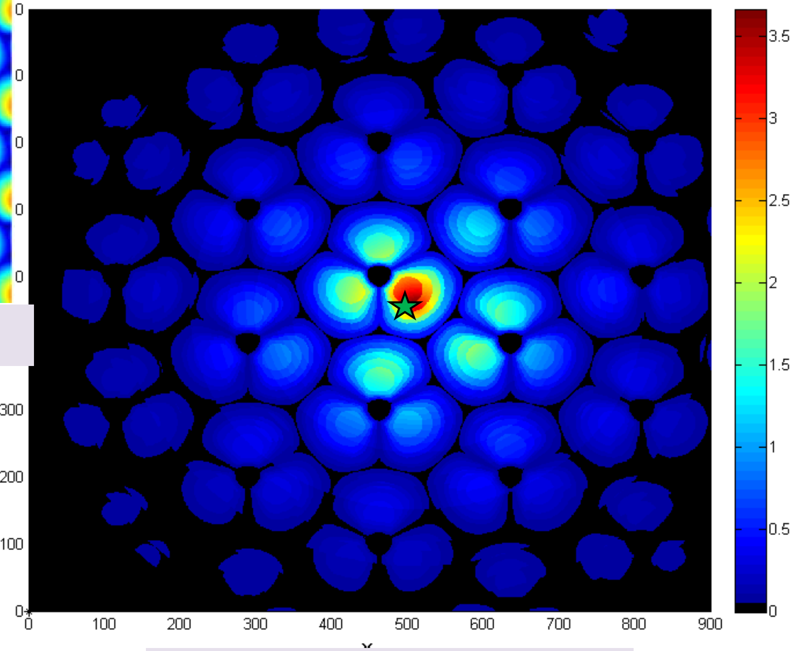
Illustration of Utility Function (Macro Cell Only)



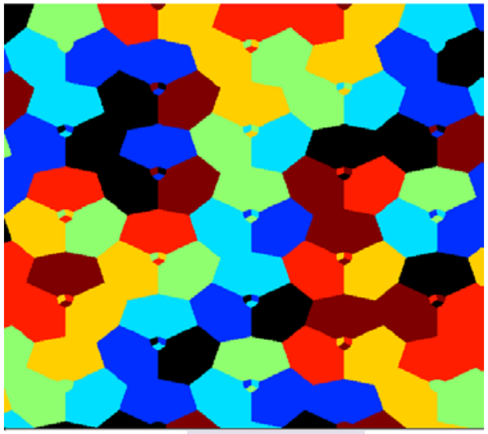
Spetral Efficiency Map max(s_j(x,y))



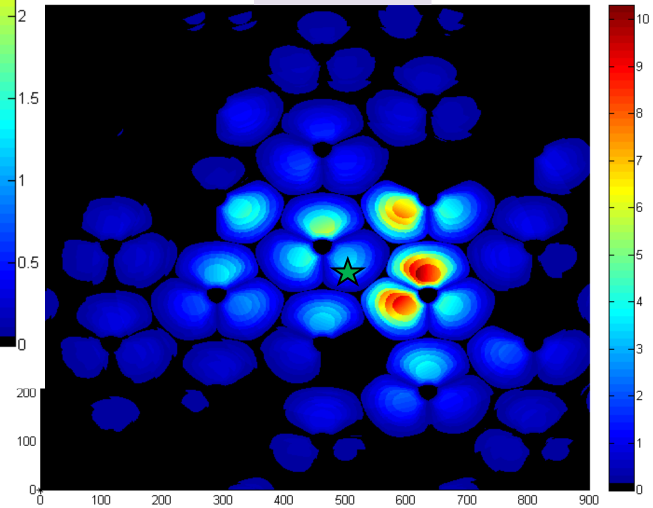
This is how p(x,y) looks like in space



p(x,y) * max (s_j(x,y))
for a typical user (green star)



Cell Load



max (Utility$_j$) for a typical user (green star)

Sequential Optimization

- Users arrive the system sequentially, and take actions independently.
- It is unrealistic to formulate the move suggestions of all users in a one-shot optimization problem.
- For a new GBR user *i*, the UIL controller conducts an exhaustive search on the utility function $U_{ij}(x,y)$ of all the cells and locations based on the current load situation.

Guarantees enough resources for user *i*

$$\begin{aligned}
\max_{j,x,y} \quad & U_{ij}(x,y) \\
\text{s.t.} \quad & W_j - \sum_{i' \in \mathcal{U}^g(i)} a_{i'j} w_{i'j} > \frac{r}{s_j(x,y)} \\
& x \in [0, X] \\
& y \in [0, Y] \\
& j \in \mathcal{C},
\end{aligned}$$

- The optimization problem of BE users are similar to that of the GBR users except that it comes without the first constraint.

Load-aware Cell Association

- Load-aware cell association is a popular approach to load balancing, however, only one type of users (in most cases, BE users) are considered in the literature.
- Using the same user model, this section is to develop a baseline load balancing approach without involving the movement of users
- Utility functions of GBR users and BE users:

$$\hat{U}_{ij} = s_{ij} \cdot (1 - \rho_j^b(i)) \qquad \hat{U}_{kj} = s_{kj} \cdot \frac{(1 - \rho_j^b(k))}{n_j^b(k) + 1}$$

- Sequential optimization of GBR user *i*:

$$\begin{aligned} \max_{j} \quad & \hat{U}_{ij} \\ \text{s.t.} \quad & W_j - \sum_{i' \in \mathcal{U}^g(i)} a_{i'j} w_{i'j} > w_{ij} \\ & j \in \mathcal{C}. \end{aligned}$$

Guarantees enough resources for user *i*

- The optimization problem of BE users are similar to that of the GBR users except that it comes without the first constrain.

Simulation Setup

Parameter	Assumption
Macrocell layout	Hexagonal grid of 19 × 3 = 57 macro-cells. Inter-site distance = 500 m
Picocell layout	57 × 2 = 114 picocells, uniformly and randomly deployed
System bandwidth	10 MHz (FDD) at 2 GHz
Average user number	10 users / cell × 171 cells = 1710 users
Percentage of GBR users	50%
Average session length	300 s
Guaranteed bit rate for GBR users	1 Mbps
Traffic model	BE and GBR
Discount incentive (δ) in UIL	-80%

The metrics are evaluated with respect to the increasing spatial traffic demand heterogeneities under three scenarios:

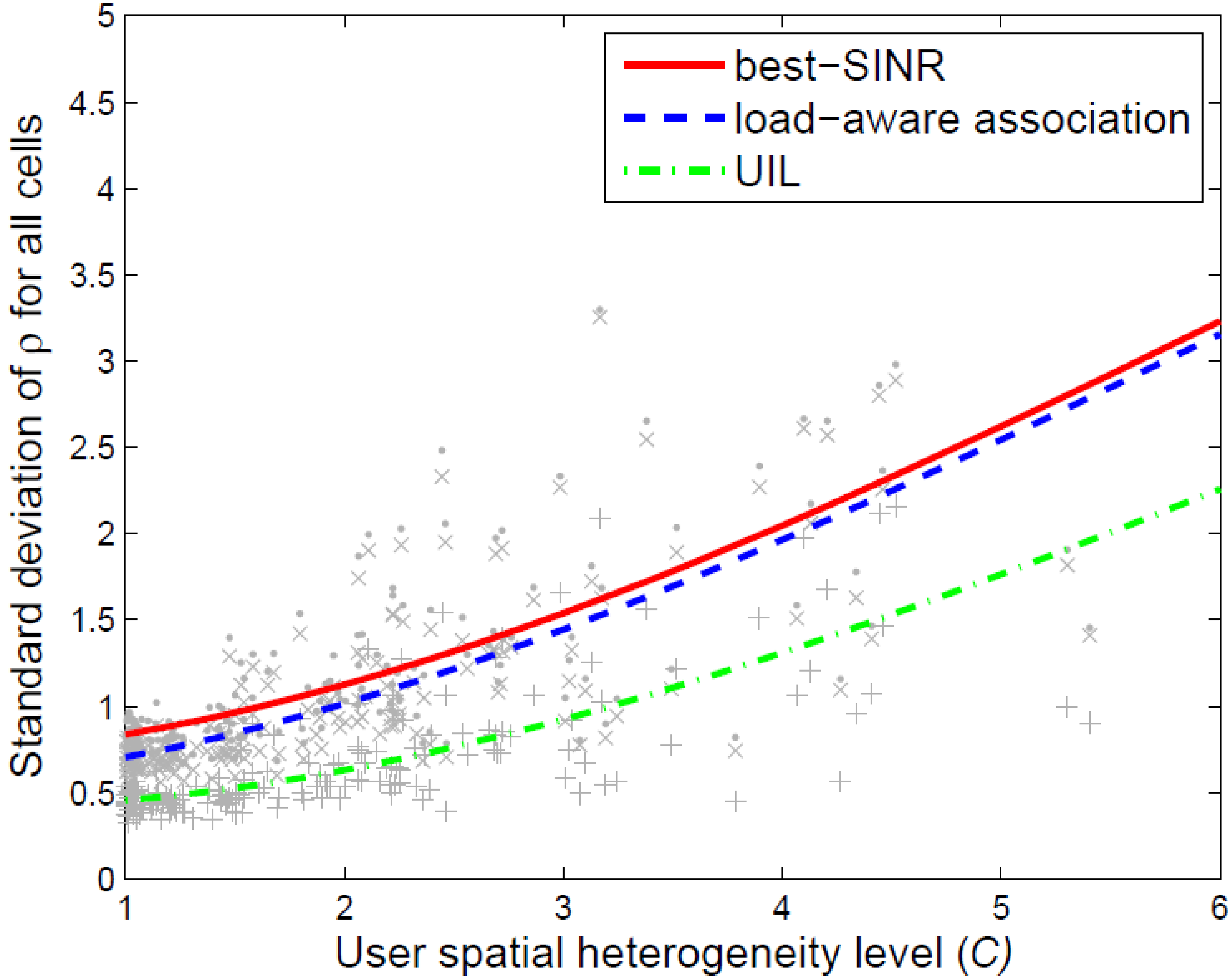
(1) no load balancing with best SINR association strategy

(2) load balancing approach with load-aware association strategy

(3) load balancing approach with the UIL scheme

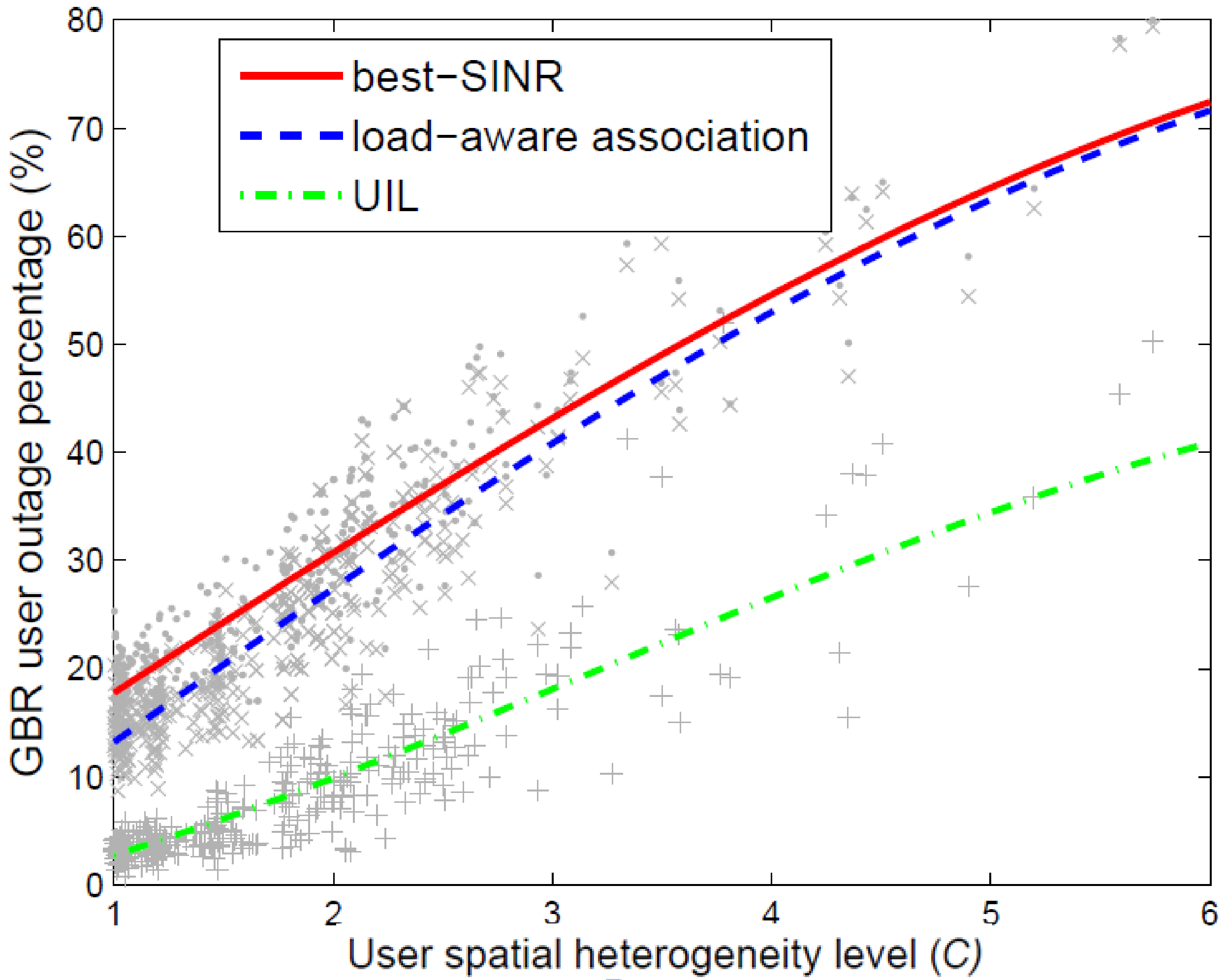
- Users arrive the system according to a Poisson process and depart the system after the session length.
- In each drop of the simulation, the system starts from zero users.
- All the performances are evaluated based on the snapshots of the system when the user number is in steady state.

Performance Evaluation – Load Balancing

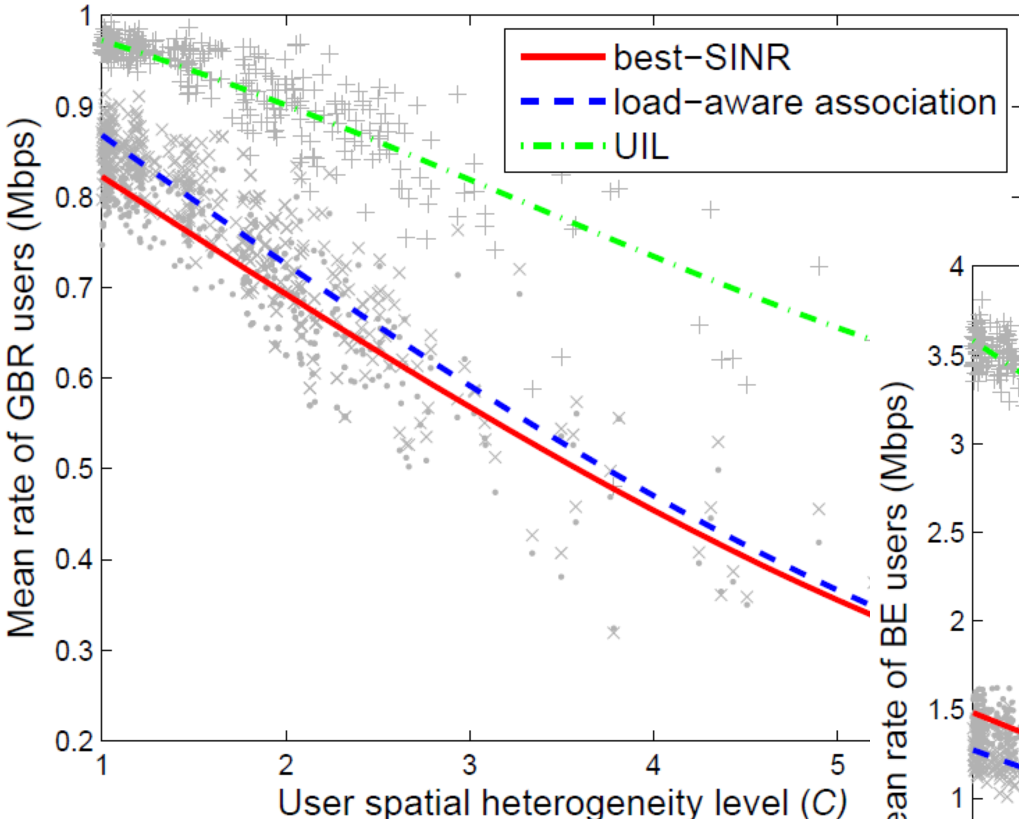


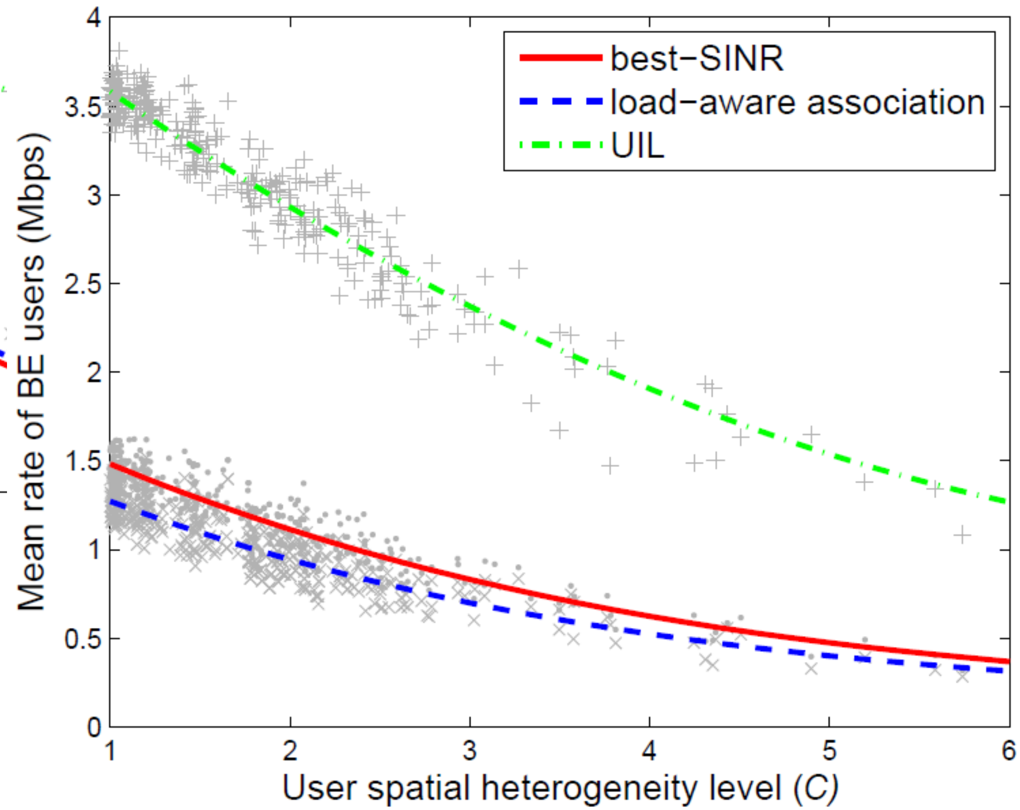
- The load of GBR users of cell *j* is defined as ρ_j in the utility function. As the GBR users have higher priority, the standard deviation of ρ_j, denoted as σ_ρ, can be used as a measure to indicate the degree of network-wide load balance.
- The lower the σ_ρ is, the more balanced the network is.

Performance Evaluation – GBR User Outage

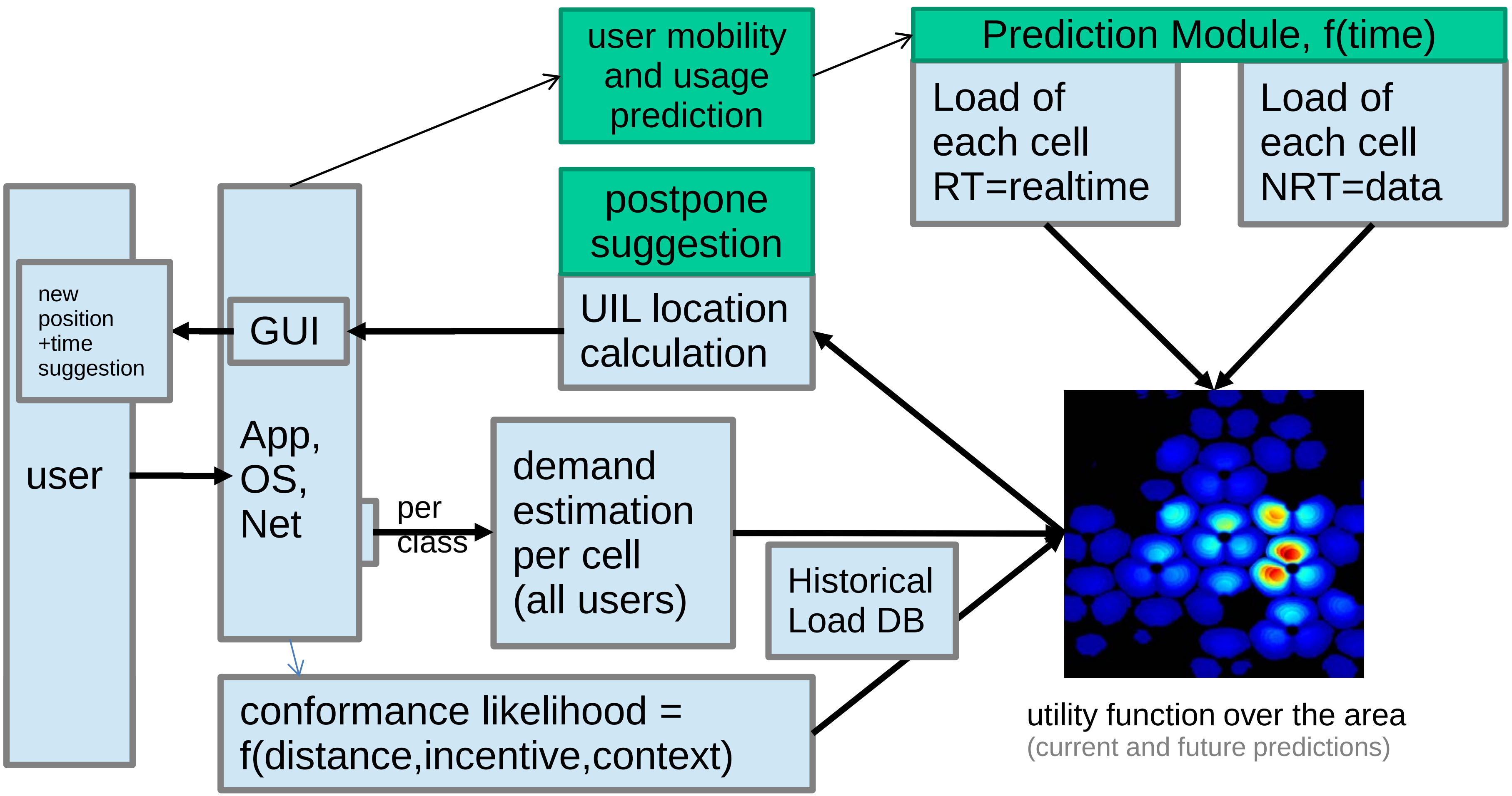


Performance Evaluation – Mean User Rate





Summary


user mobility and usage prediction
Prediction Module, f(time)
Load of each cell RT=realtime
Load of each cell NRT=data
postpone suggestion
UIL location calculation
new position +time suggestion
GUI
App, OS, Net
user
per class
demand estimation per cell (all users)
Historical Load DB
conformance likelihood = f(distance,incentive,context)
utility function over the area
(current and future predictions)

Conclusion

- A novel load balancing approach in cellular networks is proposed. User-in-loop as the spatial traffic shaping method is the enabler of the approach.
- A user model consisting of GBR user and BE user is considered in this paper with corresponding resource allocation policy.
- The intensive studied load-aware cell association approach is used as a baseline load balancing approach to compare with the traffic shaping method as proposed in this paper
- Numerical results show that the proposed load balancing approach with UIL outperforms the load balancing approach with load-aware cell association strategy and the non-load-balancing approach with best-SINR association strategy.
- The proposed load balancing approach can be extended with the temporal UIL, encouraging users to postpone a heavy data application in busy hours.